APPENDIX A: METHODOLOGY

METHODOLOGY

I. SAMPLE SELECTION

The survey drew samples of Web sites from six different target populations — all commercial U.S. sites "likely to be of interest to consumers" (group A), all such sites in the health, retail, and financial sectors (groups B, C, and D, respectively), all commercial U.S. sites "primarily directed to children aged fifteen or younger" (group E), and the most popular U.S. commercial sites (group F). The process of creating representative samples for these target populations was as follows:

1. The best available listings of sites that could be used to represent each of the six target populations were identified. These lists constituted the "sampling frames."
2. A systematic sampling procedure was used to randomly select sites from within each sampling frame. These sites constituted the "sampling pools."
3. Sites in each sampling pool were randomly examined until the number of examined sites qualifying for inclusion in the survey in each group met or exceeded the target sample sizes. The examined qualifying sites constitute the six "samples," which were surveyed for information collection practices and information practice disclosures.

The procedures used to identify sampling frames, create sampling pools, and create the final samples for the six target groups are described in the following sections. Figure A presents information about the number of sites in the sampling frame, sampling pool, and final sample for each of the six target populations.

FIGURE A

Sampling Frames, Sampling Pools, and Sample Sizes

Sample	Sampling Frame	Sampling Pool	Sample Size
Comprehensive (A)	226,644	2,408	674
Health (B)	4,140	407	137
Retail (C)	24,393	398	142
Financial (D)	6,884	398	125
Children (E)	1,483	1,483	212
Most Popular (F)	134	134	111

A. SELECTION OF SAMPLING FRAMES

1. GROUPS A-D

The Dun & Bradstreet Corporation's ("D&B") Electronic Commerce Registry database served as the starting point for creating the sampling frames for groups A through D. The database matches names and addresses registered by InterNIC, the register of domain names that is operated as a collaborative project by AT&T, Network Solutions, Inc., and the National Science Foundation, with D&B's database of 49 million businesses worldwide, and provides a comprehensive listing of commercial Web sites.[1]

As of March, 1998 there were 2,036,649 domain names registered by InterNIC in the generic or top level domains. Of these, D&B had matched over 763,499 domain names to its database of business entities in the U.S. alone.[2] Those Web sites were located in the ".com," ".net," ".org" and ".edu" generic top level domains. D&B then identified over 382,000 active commercial Web sites from those domain names. For the purpose of this study, the sampling frame for group A was restricted to 235,212 ".com" domains with active Web sites. Out of these 235,212 commercial sites, 8,568 were excluded for other reasons,[3] leaving a total of 226,644 sites in the sampling frame for group A.

Commission staff chose sites for the Health, Retail, and Financial sector sampling frames by drawing sites from the Electronic Commerce Registry based upon selected Standard Industrial

Code (SIC) classifications, which identify businesses according to their primary source of revenue.[4] This resulted in sampling frames comprising 4,140 health sector sites (group B); 24,393 retail sector sites (group C); and 6,884 financial sector sites (group D).[5]

2. GROUP E

The sampling frame for group E (children's sites) was drawn from Web sites in the Yahooligans! Directory, an online directory for children compiled by Yahoo! and located at http://www.yahooligans.com. Commission staff selected this directory because it is the largest and most diverse compilation of children's sites and provides a wide variety of commercial (".com") sites, ranging from large, corporate-sponsored sites, to small sites operated by individuals.[6] At the request of Commission staff, Yahooligans! provided a master list of 1,483 commercial U.S. Web sites in its database. This list served as the sampling frame for group E.

3. GROUP F

A sampling frame for group F (the most popular sites on the Web) was selected from three sources: Media Metrix, The PC Meter Company;[7] RelevantKnowledge;[8] and Web21's "100-Hot.com."[9] These three sources were chosen because of the diverse methodologies they employ to measure traffic at Web sites. Three separate lists of the one hundred most popular sites — one from each of the sources — served as the starting point for creating the group F sampling frame. The lists were combined into one master list, to ensure that the sampling frame represented the most popular sites on the Web. From this combined list of 300 sites, Commission staff eliminated all sites that did not have a ".com" domain name, as well as pornographic sites. Seventy sites were eliminated in this process. Duplicates were then eliminated.[10] The resulting list of 134 discrete URL's, representing any qualifying URL that appeared on at least one of the three source lists, constitutes the sampling frame for group E.[11] Finally, complete listings of all sites in the sampling frames for groups A through F were created.

B. CREATION OF SAMPLING POOLS

A systematic sampling procedure was then used to create randomly-selected sampling pools from each of the sampling frames in the following manner. First, the target sample size for each group was established. Staff set out to sample approximately 600 sites in group A, 100 sites each for groups B, C, and D, 200 sites in group E, and 100 sites in group F.

Second, a target sampling-pool size was calculated for each group. Based on a pre-test, it was estimated that approximately twenty-five percent of all sites examined in groups A through D would actually be qualifying sites for the final samples. Thus, it was estimated that four times the target sample size, or approximately 2,400 sites, would need to be examined in group A to achieve a final sample of 600 sites, and approximately 400 sites would need to be examined in groups B through D, respectively, to achieve final sample sizes of 100 sites for each group. These numbers are the target sampling-pool sizes for groups A through D. For groups E and F, the entire sampling frame was included in the sampling pool (*i.e.*, no random sampling was done of the sampling frame), resulting in target sampling-pool sizes of 1,483 and 134 respectively.[12]

Once the target sampling-pool sizes for groups A through D were determined, the actual sites for inclusion in the sampling pools were randomly selected.[13] A "sampling interval" was determined for each group by dividing the size of the sampling frame by the target sampling-pool size. For example, for group A the sampling interval was calculated to be 94 (226,644 divided by 2400, rounded off). Similarly, the sampling intervals for group B, C, and D were 10, 61, and 17, respectively.

The sampling interval was then used to randomly select sites from each sampling frame for inclusion in each group's respective sampling pool.[14] The resulting sampling pool in group A contained 2,408 sites, the sampling pool in group B contained 407 sites, the sampling pool in group C contained 398 sites, and the sampling pool in group D contained 398 sites. As noted above, the sampling pool in group E constituted all 1,483 sites in the group E sampling frame and the sampling pool in group F contained all 134 sites in the group F sampling frame.

C. FINAL SAMPLES

Once the sampling pools were created, the final samples were determined as follows. Sites in each of the six sampling pools were randomly examined until the number of qualifying sites in each group met or exceeded the target sample sizes stated in the previous section. Sites were deemed to "qualify" for inclusion in the samples if they were "likely to be of interest to consumers" (groups A-D and F) or were "primarily directed to children" (group E).[15] Non-qualifying sites were not included in the samples. Those URL's that could not be accessed for technical reasons were also not included in the samples.[16] Sites that were not included in the samples were not surveyed for information collection practices or information practice disclosures. The final sample sizes were 674 sites for the Comprehensive Sample (group A), 137 sites in the Health Sample (group B), 142 sites in the Retail Sample (group C), 125 sites in the Financial Sample (group D), 212 sites in the Children's Sample (group E), and 111 sites in the Most Popular Sample (group F).[17] Figure B shows the total number of sites examined in each group and the number of sites included in each of the final samples (final sample size), the number of sites excluded for technical reasons, and the number of sites excluded because they did not meet the qualifying definitions.

FIGURE B

Sampling Pool Disposition Table

Sample	**Sites Examined**	**Qualifying Sites**	**Non-Qualifying Sites**	**Excluded Technical**
Comprehensive (A)	1,743	674	760	309
Health (B)	223	137	46	40
Retail (C)	234	142	35	57
Financial (D)	214	125	45	44
Children (E)	1,011	212	721	78
Most Popular (F)	134	111	16	7

II. THE SURVEY

A. DATA COLLECTION

Forty Commission staff members, including attorneys, legal assistants and investigators [hereinafter "surfers"], surveyed the sites in each of the samples in the two-week period from

March 9-20, 1998. The surfers were not involved in designing the survey, or in the subsequent data analysis or drafting of this report. Each surfer underwent a full day's training in the technical skills of visiting and reviewing Web sites and in use of the survey forms.[18] Surfers conducted the survey in two survey rooms on computers using Pentium processors, minimum 90 MHz, with 32 megabytes of memory. All machines were running Windows NT 4.0 Workstation. The computers were connected to the Internet via a dedicated T1 connection. Surfers used either the Netscape Communicator 4.0 or Microsoft Internet Explorer 4.0 browser, depending on individual preference and the ease with which sites could be viewed. Computers were also equipped with the following plug-in software to maximize the number of sites that could be viewed in their entirety: Macromedia, Inc.'s Shockwave Flash 2.0 and Shockwave for Director 6.0; RealNetwork, Inc.'s RealPlayer 5.0; and ichat, Inc.'s ichat plug-in 2.22. Staff attorneys serving as supervisory proctors were present in the rooms at all times during the survey to handle any technical difficulties and answer questions.

Surfers were randomly assigned URL's (Internet addresses) from each of the sampling pools.[19] Once a URL was accessed, surfers were first required to determine whether the site qualified for inclusion in the sample — *i.e.*, whether the site was "likely to be of interest to consumers," or, in the case of the children's sample, whether the site was "primarily directed" to children aged 15 or under. A site was deemed "likely to be of interest to consumers" if it market[ed] or advertise[d] consumer goods or services **AND** it [met] one of the following two conditions:

A. it provide[d] information of interest to consumers (*e.g.*, weather, sports, stocks, research, health); or

B. it collect[ed] personal information from consumers.

To decide whether a site was "primarily directed" to children aged 15 or under, surfers were instructed to determine: whether the site used language or graphics directed to children; whether the content of the site was directed to children (*e.g.*, topics, activities, contests, pen pals, chat rooms, posting winners' home pages or art work); or whether the site collected information from children. Where surfers found that a site in the group E sampling pool appealed to an adult audience but designated an area specifically for children, the site was included in the group E sample and the survey form was completed solely with respect to the children's area.

Once a surfer concluded that a site qualified for inclusion in one of the samples, the surfer searched the site to determine whether it collects personal information from online consumers and, if so, to ascertain the kinds of information it collects, and to determine whether it discloses its information practices.[20] The data collection process is described in further detail in the body of the report.[21]

Wherever possible, surfers viewed every page of each site being surveyed. Surfers were instructed to spend up to one-half hour surveying each site and filling out its survey form. They were also instructed to print each site's home page and to print every page on which an information practice disclosure was located.

B. VALIDATION

Numerous measures were taken to ensure the quality and accuracy of the data collected by the surfers. Each site was surveyed by a second surfer who revisited the site to ensure the accuracy of the information reported on the survey form for that site.[22] Changes in answers on a site's survey form proposed by the second surfer were made only with the approval of a proctor. The completed survey forms for sites found to be posting an information practice disclosure, along with the print-outs of those disclosures, were then reviewed for errors a third time by a group of four proctors. Any suspected errors were brought to the attention of at least two proctors who jointly authorized any corrections to the survey form.

In the case of the Children's Sample, a supplemental survey form was used to review sites after the primary survey was completed.[23] A small group of surfers reviewed the initial survey forms and printouts for sites in the Children's Sample to analyze whether and how those sites addressed notice and choice for parents. In addition, the supplemental analysis required surfers to revisit these sites to distinguish between information collected from adults/parents and information collected from children, by determining the types of information collected in association with a form of payment (*e.g.*, check, money order, or credit card) which arguably required parental involvement. Information collected in any context not involving a payment form (such as contests, site or club registration, guest books, or feedback) was deemed collection of personal information from children.

Once the survey forms for all samples had undergone the multiple levels of review described above, the same data were entered into two separate databases by separate data entry personnel. The two databases were electronically compared, the survey forms of those sites with discrepancies were reviewed again, and appropriate corrections made, to ensure the accuracy of the data. A set of queries was then run on the data to ensure that the data was internally consistent, *i.e.*, that all conditional questions were answered or left blank, as appropriate. Any errors in data entry were corrected, based on the questionnaires, prior to the substantive analysis of the data.

The data analysis itself was also conducted twice, by separate individuals utilizing different analytic tools. The results of these analyses were compared to ensure uniformity and accuracy. Finally, every surveyed site was also archived to the extent possible, using Anaserve, Inc.'s Anawave Websnake 1.23 software. Due to time constraints, not all sites were archived on the same date they were surveyed; but all were archived as soon as possible after the survey was completed.

ENDNOTES

[1] The process of matching the InterNIC records to D&B's databases consists of both automated and manual quality reviews. Specifically, once the domain names are linked to businesses in the D&B database, D&B confirms that an active Web site exists behind the domain name. This confirmation is in addition to routine manual quality reviews to verify ongoing commercial activity at those Web sites.

[2] The difference between the matched domain names and the total registered domain names (a total 1,273,150 domains) represents the following: domains registered outside of the U.S. (and therefore outside the scope of this project), individuals not engaged in business operations (as defined by the D&B database), or potential matched domains that did not meet minimum requirements for match confidence.

[3] Insurance sites were excluded from the sampling frame for group A because of the restrictions concerning the "business of insurance" contained in the final proviso of Section 6 of the Federal Trade Commission Act, 15 U.S.C. § 46.

[4] Executive Office of the President, Office of Management and Budget, *Standard Industrial Classification Manual* (1987). Sites chosen for the Comprehensive Sample were drawn randomly from all SIC Codes with the exception of those for Insurance Carriers, and Insurance Agents, Brokers and Service (Codes 6311-6411). *See* note 3 *supra*.

[5] Health sector sites were selected from those sites associated with a business whose primary SIC code was 2833 (Medicinals and Botanicals), 2834 (Pharmaceutical Preparations) or in the 8011-8099 range (Health Services). Retail sector sites were selected from those sites associated with a business whose primary SIC code was in the 5211-5999 range (Retail Trade). Financial sector sites were selected from those sites associated with a business whose primary SIC code was in the 6011-6289 range (Depository Institutions, Nondepository Institutions, Security and Commodity Brokers) or in the 6712-6799 range (Holding and Other Investment Offices).

[6] Yahooligans!' selection criteria are quite broad. Yahooligans! includes commercial sites that are directed to children, as well as sites not intended for children but of interest and appropriate for children, such as sites offering information about summer camp programs or family vacations. Yahooligans! applies the criteria adopted by the American Library Association, *see* http://www.ala.org/parentspage/greatsites/criteria.html, and excludes sites that offer adult content, such as sites that promote alcohol or tobacco products. Yahooligans! surveys the Web to identify sites for its directory and also receives referrals for sites requesting to be listed. No payment is received for listing a site.

[7] Media Metrix compiles its list by using "PC Meter," tracking software installed on the computers of a representative sample of computer users. PC Meter tracks computer usage, including online usage. For more information, see http://www.mediametrix.com. The Media

Metrix list used in this report reflects online usage in January 1998.

8. RelevantKnowledge compiles its list by using a representative panel of computer users, whose online experience is then projected onto the Web universe. For more information, see http://www.relevantknowledge.com. The RelevantKnowledge list used in this report comprises findings from the period February 2 - March 1, 1998.

9. Web21 compiles its list by using logs from proxy servers placed at strategic points on the Internet throughout the world. The proxy server logs show surfing patterns for over 100,000 end-users. The survey includes university, business and home users, but does not include America Online, Prodigy or Compuserve members. For more information, see http://www.100hot.com. The "100-Hot.com" list used in this report was downloaded from www.100hot.com on March 9, 1998. The list is based on Web 21's "research" file and is different from the generic "100hot" list published by Web21 in the following respects: it is calculated based on all visits to a site (not just home-page visits) and it is based on visits to individual Uniform Resource Locators ("URL's"), as opposed to meta-sites (large Web sites that are made up of multiple servers).

10. 62 discrete URL's appeared on more than one list. As some of the sites appeared on all three lists, a total of 96 duplicate listings were eliminated from the combined master list.

11. Together, these sites represent approximately 35% of daily Internet traffic. This number is based on the sites appearing on Web21's list of 100 most popular sites. *See* http://www.100hot.com.

12. All sites were included in these sampling pools for the following reasons. Staff felt that a sampling pool approximately the size of the entire sampling frame in group E was needed to produce a sample of 200 sites primarily directed to children aged 15 or younger. The entire sampling frame in group F was used because unlike the other groups, group F was not intended to represent a random sample of sites. Rather, it reflects all of the most popular sites on the World Wide Web.

13. Because all of the sites in the sampling frames for groups E and F were included in those groups' sampling pool, no random sampling was necessary.

14. Each sampling-frame list was first numbered. For each group, a random number was then generated. The site appearing in the random-number's slot on that group's sampling-frame list was selected for inclusion in the sampling pool, as was each site appearing on the list at the interval of one sampling interval.

To illustrate, for group A, a random number table was used to choose a number between 1 and 94 (the length of the sampling interval). The number 8 was selected using this approach. Thus, the 8th site from the sampling-frame list was included in the sampling pool. Next, the sampling interval (94) was added to the random number (8) to determine the next site for inclusion, *i.e.*, the 102nd site. Repeating this procedure yielded a sampling pool of 2,400 sites that were the 8th, 102nd, 196th, 290th, etc. sites on the sampling-frame list. This process was repeated to determine sites for inclusion in the sampling pools for groups B, C and D.

15. These terms are discussed in greater detail below.

16. Sites were eliminated on technical grounds if, for example, staff found an "Under Construction," or "Inactive" message at the assigned URL, if specialized software was required to view a site, or if the URL itself could not be accessed because the URL was not recognized.

17. The following sites appear in both the Comprehensive and Retail Samples: http://www.clubmaker.com; http://haddadauto.com; and http://www.qaudio.com. In addition, the following sites also appear in more than one sample: http://www.cdnow.com appears in both the Retail and Most Popular Samples and http://www.disney.com, http://www.pathfinder.com, and http://sony.com all appear in both the Children's and Most Popular Samples.

18. Copies of the instructions and survey forms used by staff are included in Appendix B.

19. Because the survey form for children's sites was unique to that sampling pool, one group of specially-trained surfers examined only sites from the group E sampling pool.

20. As noted in Section II.B of the report, staff did not ascertain whether sites in the survey use cookies, or other hidden electronic means, to collect personal information, but instead looked to information practices disclosures to reveal such practices.

21. *See* Section V.A.

22. Findings for 99% of the sites in the Comprehensive Sample and 100% of the sites in the Health, Retail, Financial, Children's and Most Popular Samples were verified in this manner. Seven sites in the Comprehensive Sample could not be verified for technical reasons.

23. A copy of the Children's Supplemental Survey Form is included in Appendix C.